Global Digital Format Registry

Stephen L. Abrams
Harvard University
Cambridge, Massachusetts, USA

David Seaman
Digital Library Federation
Washington, DC, USA

Abstract

Detailed knowledge of digital representation formats is necessary to interpret properly the full information content of otherwise opaque content streams. The Digital Library Federation (DLF) has sponsored preliminary investigations towards establishing a Global Digital Format Registry (GDFR) that will provide persistent, unambiguous bindings between public identifiers for digital formats and the significant syntactic and semantic properties of those formats. These properties are an important component of the representation information necessary for effective digital preservation. The existence of a GDFR should prove to be of great utility to archives, libraries, digital repositories, and other organizations and individuals interested in the long-term viability of digital information.

Introduction

Effective long-term preservation of digital information requires more than merely ensuring the fixity, or data integrity, of the information bit-stream. Digital information inherently requires technological intermediation for its delivery in human-interpretable form. In order for digital representations to be rendered into an appropriate form for human consumption (an image on a screen, text on a page, etc.) the mediating software must incorporate detailed syntactic and semantic knowledge of the digital encoding. That same knowledge is necessary to perform effective digital preservation activities, regardless of whether those activities are based on an emulation or migration strategy. In both cases it is important to be able to extract the fullest possible information content from the digital objects.[1]

A digital format defines an encoding, that is, a fixed method for serializing an abstract information model into a sequence of bytes (see Figure 1). Any process that needs access to the underlying information model content, whether for purposes of rendering, interchange, or preservation, must fully understand the specific syntax and semantics of the format, or what is known in the Open Archival Information System (OAIS, recently formalized as ISO 14721) reference model as representation information.[2] Without appropriate format representation information, encoded content is reduced to an uninterpretable opaque data stream. For purposes of preservation, format representation information must be kept available over archival time-spans. (Note that the preservation process is complicated by the fact that the format representation information itself needs to be maintained in some formatted manner (PDF, XML, Word, etc.), leading to a potentially anomalous situation in which, for example, an XML specification itself encoded as XML cannot be properly interpreted until it has been properly interpreted.) The necessity for preservation intervention, whether migration or the construction of an emulator or a new delivery mechanism, will most probably occur at fairly infrequent intervals, and potentially at some remove from the time at which an appropriate format-specific technology platform was last available.

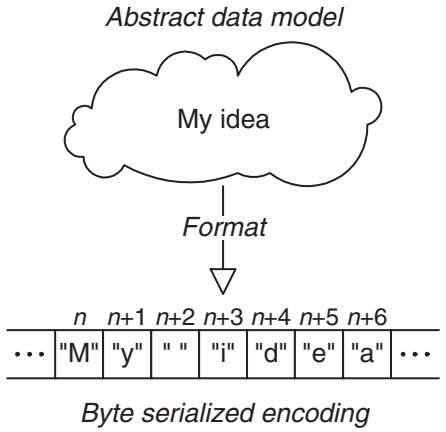

Figure 1. Digital format

It is likely that at any given point in time the overwhelming majority of objects in digital repository collections will use a relatively small number of the same digital formats. The investigation, collection, and validation

of the appropriate representation information for those formats requires extensive and specialized knowledge of the formats in question. While all repositories will need the same information, it is unlikely that they will all have the technical resources available to acquire that information locally. The existence of a common public registry responsible for the maintenance of format representation information would provide an effective mechanism to share scarce technical expertise and allow the widest dispersion of the fruits of that expertise to the appropriate community at the lowest cost to that community.[3] Such a registry would be consistent with the broad architectural guidelines of the Library of Congress's National Digital Information Infrastructure Preservation Program (NDIIPP).[4]

Descriptions of many digital formats are currently available, at varying degrees of detail and accuracy, through a variety of channels including web sites, informal reference books, and formal specification documents. Many of these sources, however, are of a transitory nature. For example, the European Commission's Information Society Technologies (IST) Programme funded the Diffuse project, which operated a high-quality web site providing extensive information on digital formats and pointers to specification documents. Unfortunately, project funding ended in 2003 and the web site was recently closed. Long-term digital preservation requires that authoritative information be available indefinitely.

The best current example of a global mechanism for authoritative format information is the MIME Media Types registry operated by the Internet Assigned Numbers Authority (IANA).[5] However, MIME registrations are text documents intended for human consumption, precluding the use of automated interactions between the registry and repositories. Furthermore, the MIME registry does not prescribe the set of format attributes that must be disclosed and under some circumstances does not require any technical disclosure. MIME types are also defined at a coarse granularity that makes no provision for families of related formats existing under a common rubric. For example, TIFF/IT (ISO 12639, used for pre-press data exchange), TIFF/EP (ISO 12234-2, output by many digital cameras), and GeoTIFF (used for geo-referenced images) are all variants of the Tagged Image File Format, but may require very different preservation processing workflows. Yet all three are identified by the same MIME type, image/tiff. These conditions make the MIME registry as currently constituted insufficient as a resource for digital preservation activities.

The digital preservation community needs a sustainable registry from which it can reliably recover authoritative format representation information defined at arbitrary levels of granularity. Towards this end the Digital Library Federation (DLF) has sponsored a preliminary investigation into the technical, operational, and business issues surrounding the establishment of a Global Digital Format Registry (GDFR). GDFR is intended to maintain persistent, unambiguous bindings between public identifiers for digital formats and representation information for those formats.

DLF-Sponsored Work

The genesis of the work towards establishing GDFR began in the summer of 2002 with informal discussions between team members of the Harvard University Library Digital Initiative (LDI) and MIT DSpace projects on topics of mutual interest. Both projects were investigating mechanisms to facilitate the format-specific aspects of digital repository operation and preservation planning. It soon became clear that the necessity for a format registry was not limited to the two projects, but rather was shared by all digital repositories and preservation programs. Under DLF sponsorship a series of invitational meetings were organized to bring together a representative group of interested stakeholders. Participation in these meetings was international in scope, including representatives from the following national and academic libraries and archives and other related organizations:

- Bibliothèque national de France
- California Digital Library
- Digital Library Federation
- Harvard University
- Internet Engineering Task Force
- Joint Information Systems Committee, UK
- JSTOR
- Library of Congress
- Massachusetts Institute of Technology
- National Archives and Records Administration
- National Archives of Canada
- New York University
- National Institute of Standards and Technology
- Online Computer Library Center
- Public Records Office, UK
- Research Libraries Group
- Stanford University
- University of Pennsylvania

In order to facilitate its analysis the working group articulated a set of potential use cases for the registry, which fell into the following categories:

- Identification – "I have a digital object; what format is it?"
- Validation – "I have an object purportedly of format *F*; is it?"
- Characterization – "I have an object of format *F*; what are its significant properties?"
- Transformation – "I have an object of format *F*, but need it in format *G*; how can I produce it?"
- Delivery – "I have an object of format *F*; how can I render it?"
- Risk assessment – "I have an object for format *F*; is it at risk of obsolescence?"

With respect to the OAIS reference model, these format dependencies exist in the repository Ingest component, which is responsible for pre-deposit validation and potential

transformation between interchange and internal format representations; the Access component, which is responsible for potential transformation between internal and delivery format representations; and the Preservation Planning component, which is responsible for obsolescence monitoring and intervention strategy definition.

Given these use cases the working group developed provisional data and service models. (For more information on this and other aspects of the GDFR project see <http://hul.harvard.edu/gdfr/>.) The data model design was informed by the OCLC/RLG white paper on preservation metadata,[6] the JISC report on its file format assessment project,[7] and the PRONOM project of the UK Public Records Office.[8] Administrative elements of the data model were suggested by ANSI X3.285, ISO/IEC 11179 and OASIS/ebXML registry standards.[9,10,11]

Data Model

The data model includes elements for the administrative properties of the registry itself as well as the various properties of the individual registered formats, including general descriptive properties such as canonical identifiers, characterization properties such as syntactic and semantic structures, processing properties such as systems and services for which registered formats are input or output, and administrative properties such as provenance. Table 1 lists the more important high-level format properties.

Table 1. High-level format properties.

Property	Type	Function
Identifier	URI	Primary or canonical identifier
Alias	URI	Variant identifier
Author	Agent	Format author
Owner	Authority	Format owner
Maintenance	Authority	Maintenance agency
Classification	Class	Ontological classification
Relationship	Relation	Arbitrary typed relationship
Specification	Document	Specification document
Signature	Signature	Internal or external signature
Tool	System	Process or service
Status	Enum	Registration status
Provenance	Event	Provenance event
Note	UTF-8	Informative note

A format can have multiple URI-based identifiers. (The exact syntax of identifiers has yet to be defined.) One globally-unique identifier, however, must be declared as the canonical identifier for the format. A format can have one or more authors, each of which can be either a personal or corporate agent. Agents are qualified by contact information and type, such as standards body, commercial business, or governmental, educational, or non-profit entity. Format intellectual property rights owners and maintenance agencies are authorities, agents associated with a specific, though possibly unbounded time-span.

All formats are given an ontological classification. The two top-level categories in the classification are Content Stream, for formats that can be usefully considered independent of media, and Physical Media, for content streams manifest only in tangible form on some physical memory structure. The Content Stream category subdivides on the basis of gross media type, while Physical Media subdivides on the basis of storage technology:

- Content Stream
  - Logical
  - Numeric
    - Scalar
      - Integer
      - Real
      - Complex
  - Text
    - Structured text
      - Mark-up
      - Programming
  - Image
    - Still
      - Font
      - Graphic
      - Page description
    - Motion
  - Audio
    - Music
  - Application
    - Communication
    - Database
    - Executable
    - Presentation
    - Spreadsheet
    - Word processing
  - Transformation
    - Compression
      - Lossless
      - Lossy
    - Container
    - Transfer
      - 7-bit safe
- Physical Media
  - Magnetic
    - Disk
    - Tape
      - Reel
      - Cartridge
  - Optical
  - Paper

The finer granularity of the classification scheme is still subject to revision.

Arbitrary typed relationships can be established between the formats in the registry to capture information such as format versioning, sub-typing (e.g., PDF/X is a PDF), and encapsulation (e.g., WAVE can contain Linear PCM). Relationships can extend to external registries to enable a distributed network of format-specific information. A central registry could potentially maintain information about formats

of broad applicability, while more specialized formats or format profiles can be stored in local institutional, regional, or consortial registries.

Formats can be associated with multiple specification documents. These are qualified by author, publisher, date, public or standard identifier (DOI, ISBN, URI, etc.), canonicity (authoritative vs. informative), and accessibility. The intent of GDFR is to include actionable links to external documents as well as maintaining local soft and hard copies of these documents if consistent with intellectual property concerns. Various levels of access to these documents will be enforced to encourage the deposit of proprietary information. These levels may include public access, on-site access only, licensed access, and escrow. All restricted access regimes will be tied to specific trigger events that will eventually bring all information into unlimited public access.

A signature is some identifying external or internal characteristic of a format such as a customary file extension, Mac OS file type, or magic number. The registry will attempt to document format-specific tools and services qualified by function and vendor contact information. All provenance events, such as initial registration, update, and delete are qualified by timestamp and agent.

Service Model

The GDFR service model provides for two categories of services: Management Services and Access Services. Management services include:

- Maintenance – Creation, update, and deletion of format registration entries
- Approval – Providing an appropriate level of technical review
- Notification – Subscription-based notification of significant events regarding specific formats
- Introspection – Machine-discoverable exposure of local registry policies and practices

Access services include:

- Description – Query mechanism for format representation information
- Export – Bulk export of registry data

A further set of ancillary services has been defined, but for the time being their implementation is being left to external value-added service providers:

- Validation – Format-specific validation of digital objects
- Rendering – Format-appropriate delivery of digital objects
- Transformation – Conversion of an object from its source format to a target format
- Metadata extraction – Metadata characterization of formatted objects

These external services would make use of the information contained within the registry to perform their functions. For example, the joint JSTOR/Harvard JHOVE project is developing a format identification, validation, and characterization framework that utilizes the same set of format representation information as is stored in GDFR and is thus a potential client of the registry.12

Prototype Registry

Based on these data and service models, a proof-of-concept prototype registry known as Fred (Format Registry Demonstrator) is under development at the University of Pennsylvania Library as part of its Andrew W. Mellon Foundation-funded project on Typed Object Models (see <http://tom.library.upenn.edu/fred/>). When completed, this prototype will be used as a testbed for refining the GDFR data and service models and suggesting an appropriate architectural design and implementation platform for a subsequent production system.

Governance Structure and Business Issues

The GDFR will be judged successful insofar as it is perceived to be sustainable and a trustworthy repository. The governance structure for the registry must be able to facilitate both of these goals. Without trust in the authoritativeness of the representation information contained within it, the registry will not be utilized by digital preservationists. Without trust in the handling of proprietary representation information, such information will not be deposited with the registry, significantly decreasing its potential value.

Sustainability of the registry is essential to provide appropriate support for long-term digital preservation. Since today's operational repositories are gracefully handling a variety of formatted material, it is often difficult to imagine how easily that community knowledge of formats can be lost with the passage of time. The GDFR will function as the persistent memory of the digital preservation community to ensure that the format knowledge we take for granted today will remain accessible to the community in the future.

It remains unclear as to whether the GDFR can operate under the administrative aegis of some existing organization or if a new organization is required. Regardless, it is important that the GDFR can be ensured of a predictable yearly revenue stream with which to fund its operation. Unlike the traditional archiving of much analog material, digital preservation is an aggressively pro-active process, requiring constant monitoring and periodic intervention to ensure the continuing viability of the material under its managed care. A momentary disruption of preservation intervention at the point of major technological change may result in the irretrievable loss of digital content. The difficulty facing the GDFR is to formulate an effective business model that will essentially provide income today for a benefit that may not accrue until tomorrow. In many ways, the administrative and business issues surrounding the GDFR will prove much more difficult to solve than the technical issues.

Next Steps

The ad-hoc GDFR working group continues to refine the data and service models, primarily through the vehicle of the Fred project. Beyond that, the group is developing a proposal to seek sufficient funding for a multi-year project to move forward towards GDFR along two tracks. The first track will involve a formal study of the governance and business issues, taken in conjunction with extensive consultation of all appropriate stakeholders, leading to specific recommendations for establishing GDFR on a firm and sustainable operational basis. The second track will build upon the lessons learned from the Fred project to design, develop, deploy, and operate a production-quality registry. The initial population of the registry will be used to explore issues and validate assumptions concerning data and service modeling, systems architecture, technical implementation, and operational considerations.

Conclusion

Effective long-term preservation of digital information will require the existence of a sustainable resource for maintaining format-specific representation information. The Digital Library Federation has sponsored an initial investigation into the technical, administrative, and business issues surrounding the establishment of a Global Digital Format Registry. An ad-hoc working group with international participation has created provisional data and service models that are being implemented in a proof-of-concept system. Funding is being sought for a multi-year two-track project that will recommend an appropriate governance and business model for an operational registry and will implement, deploy, and populate a production-quality prototype registry. It is anticipated that this project will lead to the establishment of a sustainable registry that can function as a key component of a future digital preservation infrastructure.

References

1. Julien Masanès, L'information technique nécessaire à la préservation à long terme des documents numériques [Technical information needed for long term preservation of digital documents], *International Preservation News* **29** (2003).
2. ISO 14721, *Space data and information transfer systems – Open archival information system – Reference model*, 2002.
3. Margaret Hedstrom, Seamus Ross, et al., Invest to save: report and recommendations of the NSF-DELOS working group on digital archiving and preservation (2003).
4. Library of Congress, Preserving our digital heritage: plan for National Digital Information Infrastructure Preservation Program (2003).
5. N. Freed, J. Klensin, and J. Postel, Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures, RFC 2048 (1996).
6. OCLC/RLG Working Group on Preservation Metadata, A metadata framework to support the preservation of digital objects (2002).
7. University of Leeds, Survey and assessment of sources of information on file formats and software documentation: final report (2003).
8. National Archives (UK) Digital Preservation Department, PRONOM 3 user requirements (2003).
9. ANSI X3.285, *Metamodel for the management of shareable data* (1999).
10. ISO/IEC 11179, *Information technology – Specification and standardization of data elements* (2003).
11. OASIS/ebXML, Registry information model, v2.0 (2002).
12. Stephen L. Abrams, Digital object format validation, *DLF Fall Forum* (2003).

Biographies

Stephen Abrams is the Digital Library Program Manager at the Harvard University Library, providing technical leadership for strategic planning, design, and coordination of the Library's digital systems, projects, and assets. He is currently engaged in research and implementation of effective methods for archival preservation of digital objects. Mr. Abrams is the ISO project leader and document editor for ISO/TC 171/SC 2/WG 5, the joint working group developing the PDF/A standard. He is a member of ACM, ALA, ASIS&T, and the IEEE Computer Society.

David Seaman is the Executive Director of the Digital Library Federation. Mr. Seaman was formerly director of the Electronic Text Center of the University of Virginia Library.